scGPT: toward building a foundation model for single-cell multi-omics using generative AI

Year: 2024 | *Citations: 629* | *Authors: Haotian Cui, Chloe Wang, Hassaan Maan*

Abstract

Generative pretrained models have achieved remarkable success in various domains such as language and computer vision. Specifically, the combination of large-scale diverse datasets and pretrained transformers has emerged as a promising approach for developing foundation models. Drawing parallels between language and cellular biology (in which texts comprise words; similarly, cells are defined by genes), our study probes the applicability of foundation models to advance cellular biology and genetic research. Using burgeoning single-cell sequencing data, we have constructed a foundation model for single-cell biology, scGPT, based on a generative pretrained transformer across a repository of over 33 million cells. Our findings illustrate that scGPT effectively distills critical biological insights concerning genes and cells. Through further adaptation of transfer learning, scGPT can be optimized to achieve superior performance across diverse downstream applications. This includes tasks such as cell type annotation, multi-batch integration, multi-omic integration, perturbation response prediction and gene network inference. Pretrained using over 33 million single-cell RNA-sequencing profiles, scGPT is a foundation model facilitating a broad spectrum of downstream single-cell analysis tasks by transfer learning.